

Conference Abstract

Biodiversity Literature after Disentis: FAIR, AI-assisted, XML-First Publishing Workflows

Laurence Benichou[‡], Lyubomir Penev^{§,|}[‡] MNHN, Paris, France[§] Pensoft Publishers, Sofia, Bulgaria

| Institute of Biodiversity & Ecosystem Research - Bulgarian Academy of Sciences, Sofia, Bulgaria

Corresponding author: Laurence Benichou (laurence.benichou@mnhn.fr)

Received: 23 Dec 2025 | Published: 24 Dec 2025

Citation: Benichou L, Penev L (2025) Biodiversity Literature after Disentis: FAIR, AI-assisted, XML-First Publishing Workflows. Biodiversity Information Science and Standards 9: e183376.

<https://doi.org/10.3897/biss.9.183376>

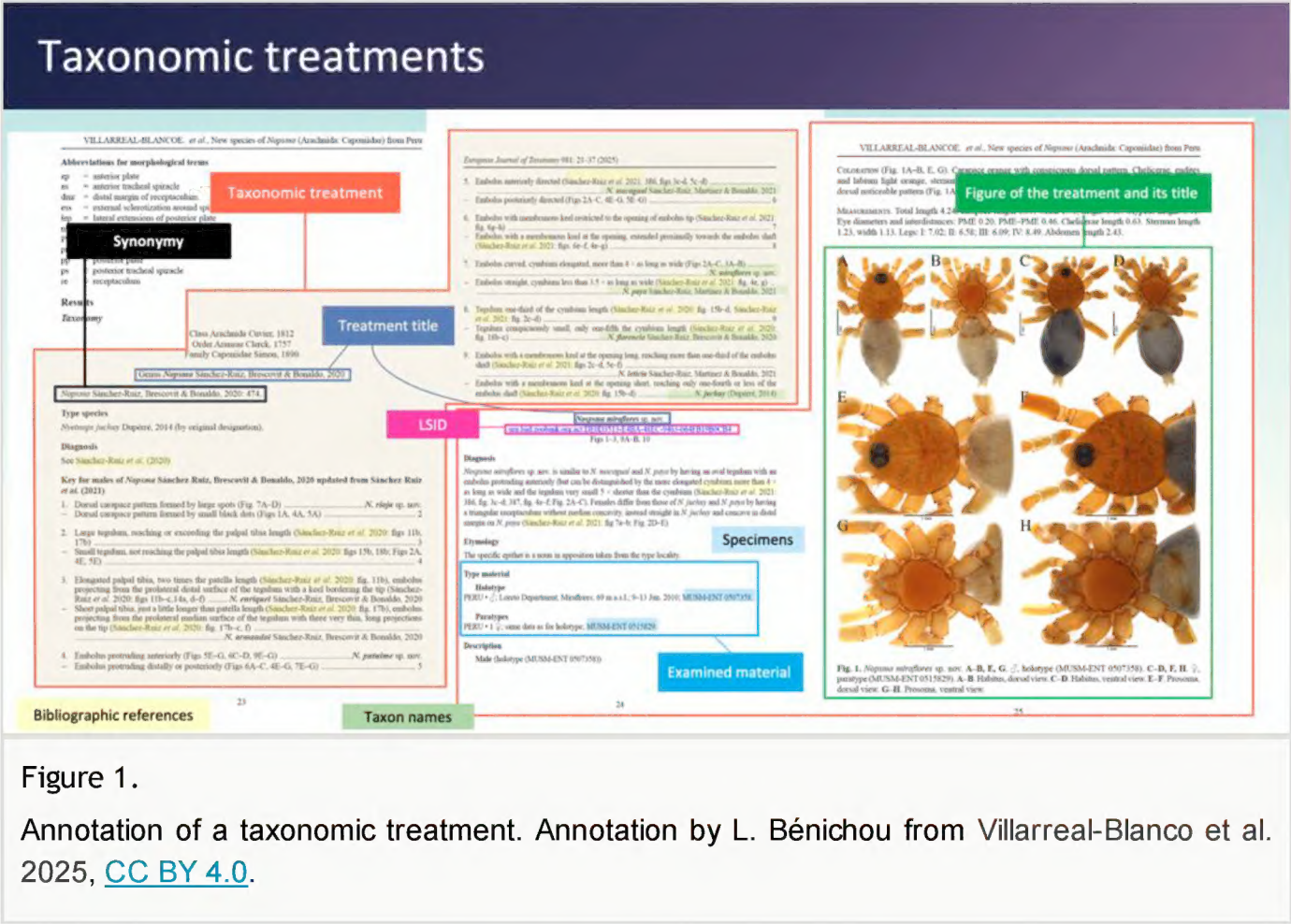
Abstract

For over a decade, a lot of effort has been put into making millions of pages and taxonomic treatments available online. Yet, too often, biodiversity knowledge remains locked in hardly accessible and non-machine-readable literature and electronic resources. Indeed, most prospective publications are still produced in an “old-fashioned” way (PDF), thereby posing a risk of following a pattern we have seen with legacy literature from the past centuries.

Even though access barriers can be demolished using retroconversion (i.e., converting a PDF into an XML annotated file once the article is published), it is far more efficient to have author- and editor-vetted annotations and semantic enhancements to the texts and data ready ahead of publication. Doing so avoids discrepancies, and ensures earlier and rapid dissemination and re-use of data. Beyond open access, the challenge for publishers is to reach interoperability, which would ultimately shorten the too-long time to find and analyse past literature (Fontaine et al. 2012) and therefore speed up the description of taxonomy.

Making data within publications findable, accessible, interoperable, and reusable ([FAIR](#)) implies text structuration, semantic annotations*¹, and standardization. Furthermore, to link data, information and knowledge contained in literature and other electronic

resources to uniquely identifiable components (i.e., images, tables, references, taxonomic treatments; see Fig. 1) means that those would be re-usable in research and policy covering biodiversity and other domains. Ultimately, this enables immediate re-use of the data, and the integration of publications into a comprehensive global biodiversity knowledge graph.



Following the vision of the [Disentis roadmap](#)^{*2}, this presentation will outline and demonstrate the value of advanced methods for scholarly publishing, production and FAIRisation of biodiversity data during the journal production process. We described the latest developments in the scientific publishing industry and—more specifically—how to publish linked and semantically enhanced research outcomes. The presentation clarifies the concepts that are relevant within the publication for semantic enhancement. Highly-automated, XML-first, AI-assisted existing workflows for semantic enhancement of articles are introduced, including the related standards, all dedicated to making published data FAIR by creating bi-directional linked data from and to the published article.

Semantization goes beyond metadata and all relevant information for biodiversity are isolated, annotated, and linked with persistent identifiers (Chester et al. 2019, Agosti et al. 2022) including: taxonomic treatment, taxon names, material examined, and material citation, the latter including country, locality, coordinates, collection date, collectors, and specimen code.

Once annotated, the XML-based publication can be disseminated through a partnership with [Plazi](#) and all components pushed to relevant databases, e.g., [Ocellus](#), Global Biodiversity Information Facility ([GBIF](#)), [Biodiversity PMC](#) (PubMed Central), [Checklist](#)

[Bank](#), National Center for Biotechnology Information ([NCBI](#)). The potential reuse of data is multiple and will benefit the publishers as well as the research community. Indeed, this wide dissemination of data contained within their articles enables publishers to create dashboards to measure e.g., numbers of articles published, numbers for data described, the richness of the data, the number of occurrences by continent.

The benefits are multiple: not only do XML-first journal production workflows save time, effort and provide FAIR-born data on the day of publication, they optimize and accelerate the dissemination of data while providing human- and machine-readable publications. Semantically structured FAIR data pave the way for training and use of AI in publishing.

Keywords

FAIR data, biodiversity literature, scientific annotation, XML

Presenting author

Laurence Benichou

Presented at

Living Data 2025

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Agosti D, Benichou L, Addink W, Arvanitidis C, Catapano T, Cochrane G, Dillen M, Döring M, Georgiev T, Gérard I, Groom Q, Kishor P, Kroh A, Kvaček J, Mergen P, Mietchen D, Pauperio J, Sautter G, Penev L (2022) Recommendations for use of annotations and persistent identifiers in taxonomy and biodiversity publishing. Research Ideas and Outcomes 8 <https://doi.org/10.3897/rio.8.e97374>
- Chester C, Agosti D, Sautter G, Catapano T, Martens K, Gérard I, Bénichou L (2019) EJT editorial standard for the semantic enhancement of specimen data in taxonomy literature. European Journal of Taxonomy 586 <https://doi.org/10.5852/ejt.2019.586>
- Fontaine B, Perrard A, Bouchet P (2012) 21 years of shelf life between discovery and description of new species. Current Biology 22 (22). <https://doi.org/10.1016/j.cub.2012.10.029>
- Villarreal-Blanco E, Martínez L, Eyes-Escalante M (2025) Four new species of *Nopsma* Sánchez-Ruiz, Brescovit & Bonaldo, 2020 (Arachnida: Caponiidae) from Peru. European Journal of Taxonomy 981 (1): 21-37. <https://doi.org/10.5852/ejt.2025.981.2807>

Endnotes

- *1 Semantic annotation is the process of assigning relevant metadata information to concepts and their relationships in a document. It enriches the document with meaningful data by linking ontology-based metadata, providing a unified structure for data integration. Semantic annotation enables interpretation by both humans and machines, reducing ambiguity and increasing reusability. It supports data analytics, intelligent querying, and reasoning. Definition from [Science Direct](#).
- *2 The [Disentis Roadmap](#) is a decadal strategy for liberating global biodiversity knowledge from scientific literature. It emerged from a symposium held in Disentis, Switzerland, in August 2024, which assessed the impact of the 2014 [Bouchout Declaration on Open Biodiversity Knowledge Management](#) and identified strategic priorities for the coming decade. As of November 13, 2025, the Roadmap has been signed by 107 entities.